DOCUMENT RESUME

ED 300 381 TM 012 169

TITLE Illinois Initiatives for Education Reform. Test Preparation Program for Gifted and Talented Sophomores: 1986 Summer Program. Evaluation Report.

INSTITUTION Chicago Board of Education, Ill. Dept. of Research and Evaluation.

PUB DATE 87

NOTE 38p.; This document is printed on blue paper.

PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Academically Gifted; *Achievement Gains; College Entrance Examinations; Grade 10; High Schools; *High School Students; *Program Evaluation; Scores; Student Attitudes; *Summer Programs; Talent; Teacher Attitudes; *Test Coaching; Test Wiseness

IDENTIFIERS *National Merit Scholarship Qualifying Test; *Preliminary Scholastic Aptitude Test

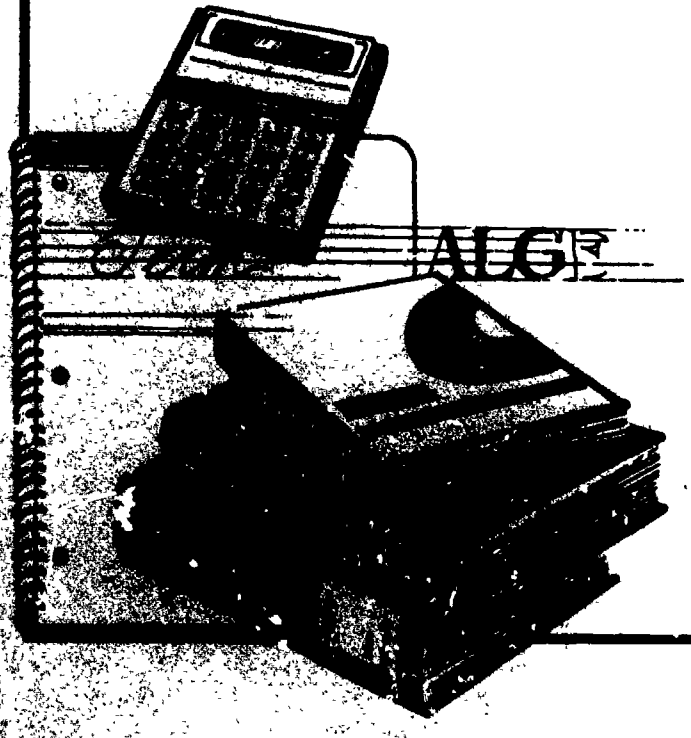
ABSTRACT

The fourth year of the Test Preparation Program for Gifted and Talented Sophomores (TPPGTS) is evaluated. This 6-week, 75-hour test coaching program was developed to teach high-achieving students principles/strategies required for doing well on the Preliminary Scholastic Aptitude Test (PSAT)/National Merit Scholarship Qualifying Test to increase the number of National Merit semifinalists and finalists. The TPPGTS emphasized language arts, mathematics, and guidance, and was conducted at the Lane, Lindblom, and Curie High Schools in Chicago, Illinois. Pretests and posttests were completed by 93 of the 148 enrollees; 72 of these took the October PSAT. The comparison group initially included 137; of these, 47 completed practice tests and 37 took the PSAT. There were nine teachers and three teacher aides in the TPPGTS. Student and teacher questionnaires and classroom observations were analyzed concerning the degree to which summer program students outperformed the comparison group from pre- to posttest and whether this effect was attributable to the program. The TPPGTS appeared to be implemented as designed; an overall positive math program effect equivalent to 42 SAT points was found. Neither general nor site-specific verbal coaching effects were seen. Students were generally satisfied with the TPPGTS and thought they had learned effective strategies. Teachers noted the diversity of instructional materials and activities, and made recommendations for improving the program. (SLD)

Test Preparation Program for Gifted and Talented Sophomores

1986 Summer Program
EVALUATION REPORT

Illinois Initiatives for Education Reform



Chicago Public Schools
Manford Byrd, Jr.
General Superintendent of Schools

DEPARTMENT OF RESEARCH AND EVALUATION
BUREAU OF PROGRAM EVALUATION

ILLINOIS INITIATIVES FOR EDUCATION REFORM

TEST PREPARATION PROGRAM FOR GIFTED AND TALENTED SOPHOMORES

1986 Summer Program
Evaluation Report

Chicago Public Schools
Manford Byrd, Jr.
General Superintendent of Schools

It is the policy of the Board of Education of the City of Chicago not to discriminate on the basis of race, color, creed, national origin, religion, age, handicaps unrelated to ability, or sex in its educational program or employment policies or practices.

Acknowledgments

The program evaluated in this report was funded under the Illinois Initiatives for Education Reform. Program development and implementation was jointly administered by the Office of Programs for Gifted and Talented Students and the Bureau of Guidance Programs and Services.

The evaluation was designed and conducted by the Bureau of Program Evaluation, Department of Research and Evaluation, and coordinated by Geraldine L. Oberman. The evaluation specialist was Arthur Reynolds. Staff from the Department of Research and Evaluation provided clerical, technical, and supportive assistance.

Table of Contents

Executive Summary

This report examines the fourth year (1986) of the Test Preparation program for Gifted and Talented Sophomores (TPPGTS). The TPPGTS is a six-week, 75-hour coaching program developed to teach high achieving students test principles and strategies necessary to do well on the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT). The test is administered every year in October. The central goal of the program is to produce more National Merit semi-finalists and finalists than would be expected without such a program. The program was conducted at three sites, Lane, Lindblom, and Curie High Schools, and ran from June 27, 1986, to July 31, 1986. Students were recruited citywide. Ninety-three (93) out of 148 students (63%) completed practice pre- and posttests during the program while 72 summer program students took the October 1986 PSAT. One-hundred thirty-seven comparison group students were initially identified and 47 (34%) completed practice pre- and posttests. Of this group, 37 took the October 1986 PSAT.

The major evaluation question addressed was the degree to which summer program students outperformed the comparison group from pre- to-posttest and whether this effect was attributable to the program. Also of interest was the effectiveness of program implementation and student and teacher perceptions of the program. The primary results of the evaluation are as follows:

A. The program appeared to be implemented as designed and students were on-task during observation periods. A variety of instruction materials and activities were employed, although the diversity of instructional approaches may have compromised program uniformity across sites.

B Results of pre- and posttest PSAT scores indicated a substantial math program effect at Lane as summer program students outgained their comparison group by 7.1 points. Results were more modest at Lindblom (summer program gain of 3.3 points) and Curie (1.4 points). An overall positive math program effect of 4.2 points (42 SAT points) was found. However, general or site-specific verbal coaching effects were not found as comparison group students slightly outperformed summer program students (1.9 to 1.5 point gains).

C. Stanine group analysis did not support higher PSAT gain scores for higher stanine groups. Actual score differences between stanine groups indicated that average seventh and eighth stanine students are least likely to become National Merit semi-finalists.

D. Students were generally satisfied with TPPGTS as they indicated they became better prepared to take the PSAT, thought the materials were effective, and learned a variety of test-taking strategies. However, despite their training, one-third of the program students indicated they would not usually guess if they didn't know an answer to a problem. This is a large percentage given the emphasis that is placed on making educated guesses.

E. Teachers noted the diversity of instruction materials and activities used in the program. Primary activities included discussion, independent seatwork, oral recitation, and demonstration. The teachers also made program recommendations including improving the criteria of selection, changing the program to after-school, and providing more teacher in-service training sessions.

Thus, the major outcome of this study was the substantial impact of the math coaching program, especially at Lane. This effect was traced to the systematic use of former PSATs during the program and was critically analyzed for internal and external validity. These results are more positive than other math coaching studies. The verbal coaching program results were a disappointment but are not supported by results of the 1985 and 1984 programs. An emphasis on practice and drill with former PSATs is suggested. Future programs should also mandate uniform program implementation to duplicate program effects across sites. The following recommendations are also made:

1. Emphasize practice and drill with past PSATs.

2. Train teachers about the most effective ways to coach the test.

3. Standardize the instruction and materials.

4. Improve the student selection process.

5. Refine the identification and selection of gifted students to be served by this program.

6. Change the program to a general SAT/ACT preparation program.

7. Give students an incentive for participating in the program.

8. Since coaching programs produce short-term gains and do not improve cognitive skills necessary for doing well in college, emphasis should be given to programs that develop long-term cognitive skills.

Introduction

This report examines the fourth year (1986) of the Test Preparation Program for Gifted and Talented Sophomores (TPPGTS). The TPPGTS is a six-week intensive coaching program developed to teach high achieving students test principles and strategies necessary to do well on the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT). The test is administered every year in October and is composed of verbal and quantitative items deemed relevant for college preparation. The central goal of the program is to produce more National Merit semi-finalists and finalists than would be expected without such a coaching program. The program ran from June 27, 1986, to July 31, 1986, and was approved and authorized by the Board of Education on April 30, 1986. The program recruited students citywide and was located at three sites: Lane, Lindblom, and Curie high schools.

Research Perspective

Test coaching has had a short but controversial history in education. During the 1950's, early studies of the effects of practice and coaching on test performance were conducted in Great Britian and dealt with tests used to assign children to different secondary schools (Anastasi, 1981). It was typically found that the degree of improvement depended on ability, educational experience, type of coaching, and the underlying characteristics of the test. Coaching programs for the Scholastic Aptitude Test (SAT) were also being initiated at this time. Early studies showed promising results despite pronouncements by the Educational Testing Service (ETS) and College Entrance Examination Board (CEEB) that such coaching was ineffective.

Since these early developments there has been extensive research on the effects of coaching on test performance, especially regarding the SAT. Dyer (1953), of ETS, conducted one of the first SAT coaching studies and found that program students made significant mean gains over the control group of 12.9 points on the math subtest and 4.6 points on the verbal subtest. The 10-hour program was composed of drill and vocabulary development on items developed by ETS to be compatible with the SAT. Other studies of relatively short coaching programs conducted by ETS (Dear, 1958; French, 1955) showed even higher significant gain scores for SAT-coached students.

To synthesize the plethora of SAT and PSAT coaching studies, two extensive literature reviews (Messick & Jungeblut, 1981; Slack & Porter, 1980) and a meta-analysis (DerSimonian & Laird, 1983) have been completed. Slack & Porter (1980) examined the results of 19 published SAT coaching studies, including those mentioned above, and found average gain scores of 29 and 33 points on the verbal and math subtests, respectively, although many of these studies did not include control groups.

Messick & Jungeblut (1981) confirmed the significant positive effects of SAT coaching studies but indicated that observed gains were reduced for controlled or randomized designs. They also found that program effects increased with student contact time. Using a logarithmic function analysis,

they computed average score effects as related to the number of contact hours for the 17 verbal and 14 math studies. From this model it was estimated that a verbal gain of 10 points would require 12 hours of instruction and a verbal gain of 20 points would require 57 hours. For the math subtest, 20 score points would require 19 hours of instruction, 30 score points would require 45 hours of instruction, and 40 points would require 107 hours. DerSimonian & Laird (1983), in a meta-analysis of published SAT test preparation programs, found that program effects differed by the type of evaluation model employed. Using many of the same studies as the above reviews, they found average program effects to be approximately 10 points on the verbal and math subtests for matched and randomized studies.

Thus, despite early pronouncements by ETS that special coaching can do little to improve SAT scores, research has consistently shown that statistically significant gains due to coaching can be expected from typical test preparation programs and gains generally increase with student contact hours. Based on the above studies, estimated average SAT coaching gains appear to be 10-20 points on the verbal and math subtest, although the verbal subtest is less amenable to coaching than the math subtest. However, it should be remembered that the studies cited varied considerably in the kind of instruction/materials used and may not represent the best available means to producing maximum coaching gains. For example, the recent availability of past SAT forms provides more advantageous means for practice and drill on SAT items. The few randomized and/or highly controlled studies actually completed and their short duration further limit the above results. It is imperative that studies be undertaken with more diverse populations and instructional approaches before the magnitude of coaching effects is confirmed. Most reported studies include graduating seniors in private high schools and are not generalizable to all student groups such as gifted students, for example.

Results of the past two years (1984-1985) of TPPGTS indicated average gains of 5.6 and 3.8 points on the verbal and mathematics subtests, equivalent to 56 and 38 points on the SAT scale (200-800 range) (Chicago Public Schools, 1986). While these results are much greater than results of the above meta-analyses, they are tempered in two ways: (1) the evaluations did not employ adequate control groups designs, thus probably overestimating program effects, and (2) the length of the programs (75 hours) and the type of students enrolled (gifted) would be expected to produce higher gain scores than typical test preparation programs.

The present evaluation of the fourth year of the TPPGTS takes into account the major limitations of previous years by employing a pretested control group design to determine program effects. Program and comparison group students were tested at the beginning and end of the summer program with past forms of the PSAT. Classroom observations and questionnaires were also employed to assess the implementation of, and participant satisfaction with, the TPPGTS. In addition, results of the October 1986 PSAT were obtained and subsequently compared across groups for stability. The use of October 1986 PSAT scores in this way differs from previous evaluations, in which program effects were related to the October test. With the addition of comparison group pretests, it is no longer advantageous to compare October test results for determination of program effects.

Evaluation Questions

The following evaluation questions were addressed in this study to measure program effectiveness:

A. Was the program implemented as designed? Were the instruction materials compatible with this design?

B. Did students participating in the summer program show greater improvement than comparison group students from pre- to posttest and was this improvement stable up to the October 1986 PSAT?

C. What was the students' assessment of the program and its materials? What did they learn from the program?

D. What were the perceptions of the teachers in the program regarding its content? What perceptions did teachers attribute to students?

These questions will be of particular interest to groups who have a vested interest in this program including (1) students interested in preparing for standardized tests, (2) teachers devoted to test preparation instruction and facilitating testwiseness among their students, (3) counselors who are most interested in assisting students with academic and career plans, (4) school administrators in general who are crucial in communicating the relevance of test preparation to students, and (5) the educational and scientific community at large who can help facilitate and develop the evaluation of test preparation programs.

Evaluation Design

In order to test the central evaluation question, whether summer program students show greater growth on pre- and posttest PSAT tests than comparison group students, an untreated control group design was used (Cook & Campbell, 1979). This generally interpretable nonequivalent control group design tests criterion score differences between groups before and after the program. Determination of program effects is estimated to be the difference between mean criterion score growth of experimental and comparison groups, that is, the mean score differences from pre- to posttest between summer program and comparison group students on PSAT verbal and mathematics subtests. To determine the stability of program gain scores, results of the October 1986 PSAT were obtained and analyzed for fitness. While this evaluation design is generally interpretable, it is less desirable than a control group design with controlled selection whereby students are randomly assigned to treatment conditions. Thus, selection differences between groups can bias results. The ethical question of whether to exclude students from participation in a program with documented positive effects precluded employment of this more powerful design.

Student Information

One hundred forty-eight (148) students enrolled in TPPGTS at three sites and were obtained from over 1400 students eligible to participate. Eligibility, as defined by program administrators, was restricted to students

who scored at or between the seventh and ninth stanines in reading and math on the Fall 1985 Tests of Achievement and Proficiency (TAP). Fifty-seven (57) percent of program students initially attended the program at Lane while 19 (28) and 24% (35) attended at Lindblom and Curie, respectively. Sixty-three (63) percent or 93 of the participating students completed the program and pre-posttests.

Demographic information obtained from students completing both pre- and posttests indicated that 51% were female, 49% were male, and 79% will be in their third year of high school as of September 1986. Twenty-one (21) percent will be sophomore high school students. Nearly all program students (94%) plan to take the October 1986 PSAT and 44% took the PSAT prior to the program. Eleven (11) percent of the students have been involved in other test preparation programs. In addition, 22 schools were represented in the program as 30% of enrolled students regularly attended Lane; 11% attended Curie; 10% attended Young, and 10% attended Lincoln Park. Kenwood and Von Steuben students each accounted for 9% of program students. See supplemental report documentation for a more detailed account of evaluation instrument responses.

Regarding the nature of their enrollment, 76% of the students reported registering for the program because they wanted to learn to do better on tests. Sixty-four (64) percent of the students indicated they enrolled because they thought the program would be interesting. Other reasons cited by students for participating in the program were they had nothing else planned for the summer (32%), their parents expected them to (30%), and they needed the English review (29%). It should also be noted that many eligible students were not sent letters of participation because of errors in mailing. While some of these students were finally contacted through notices sent to potential comparison group students, it is likely these letters did not compensate for those not sent.

The 137 comparison group students were also obtained from the pool of eligible students through letters sent independently of the summer program. However, these letters were restricted to those eligible students who attended Lane, Lindblom, Curie, Young, and Kenwood. Comparison group students, who may not have had time to participate in the program, responded to letters of invitation to take two practice PSAT's before and after the summer program at Lane and Curie in preparation for the October 1986 PSAT. Thirty-four percent (47) of the students tested comoleted pre- and posttests. As with the summer program group, a majority of students responding to an information sheet were female (53%) and were in their third year of high school (82%) as of September 1986. In addition, all students indicated they were planning to take the October 1986 PSAT while 22% of responding students had taken the PSAT before. Twenty (26) percent of the comparison group students also indicated they will study for the October 1986 test. Further, over one-half of the comparison group students (65%) regularly attended Lane and 26% attended Young. Seven and two percent of participating students attended Kenwood and Lindblom, respectively.

Also notable were the major reasons comparison group students did not enroll in the summer program. These included vacation, illness, or a summer job (54% of students), willingness to study for the PSAT on their own (26%), summer courses conflicted with the program (26%), and letters informing them

about the program arrived too late (15%). An additional 24% of the students indicated other reasons for not enrolling in the program, most frequent of which was they never received notification of the program. This finding is reflective of the fact that letters were not sent to all eligible students.

TABLE 1

Distribution of TAP Stanine Scores
by Summer Program and Comparison Groups
(Summer Program N=93, Comparison Group N=47)

	Reading			Math	
Stanine	Summer Program	Comparison Group	Stanine	Summer Program	Comparison Group
7	33%	34%	7	40%	34%
8	40%	32%	8	36%	43%
9	27%	34%	9	24%	23%
Total	100%	100%	Total	100%	100%
Mean ($\bar{X}$)	7.9	8.0	Mean ($\bar{X}$)	7.8	7.9

Table 1 displays fall 1985 TAP stanine scores by summer program and comparison group students who completed both pre- and posttests. As is shown, mean reading (7.9 and 8.0) and math (7.8 and 7.9) stanine scores are approximately equal between summer program and comparison group students, respectively. Thus, the academic achievement of both groups on these measures is equivalent. The distribution of TAP stanine scores between groups is also similar. On the reading subtest, comparison group students were evenly divided among stanine groups, while stanine eight summer program students had a slightly higher proportion of students (40%) than the comparison group. On the math subtest, differences between groups occurred at the seventh and eighth stanines. The highest proportion of summer program students was at the seventh stanine (40%), the reverse of the comparison group, in which 43% were at the eighth stanine. It should also be noted that TAP subtest and total test scaled scores were also compared between groups. Results were similar to the above stanine data as no significant differences existed.

Table 2

Self-Reported High School English and Math Records by Group
(Summer Program N=87, Comparison Group N=46)

ENGLISH COURSES					MATH COURSES				
	Summer Program		Comparison			Summer Program		Comparison	
Class	% Enrolled	Mean Grade	% Enrolled	Mean Grade	Class	% Enrolled	Mean Grade	% Enrolled	Mean Grade
English 1	96	3.19	100	2.98	8th grade algebra	46	3.19	42	3.13
English 2	94	3.04	100	2.95	Algebra I	97	3.20	98	3.10
English 3	11	2.50	11	3.00	Geometry	92	3.01	98	2.98
					Algebra II	28	3.35	22	2.67
Mean # Courses Enrolled/ Grade	2.1	3.10	2.1	2.96	Trigonometry	32	2.96	24	3.70
					Mean # Courses Enrolled/Grade	3.0	3.09	2.9	2.94

Further academic information between groups is reported in Table 2. As displayed, the proportion of summer program and comparison students having taken basic English and math courses is nearly equivalent across all courses. Over 90% of both groups have taken English 1, English 2, Algebra 1, and Geometry. In addition, self-reported math and English grades between groups were similar as average course grades indicate. The only noticeable difference between groups occurred in Algebra II where the mean grade of program students (3.35 out of 4.00) was higher than comparison group students (2.67). However, the small sample size of the comparison group and instructional differences between classes may account for this difference. Thus, the similarity of high school courses taken and grades received for the groups strongly indicate the general academic equivalence of both groups.

Teacher Selection

Nine teachers were assigned to the PSAT preparation program, three at each of the three sites. Teachers at each site specialized in the content areas of language arts, matheratics, or guidance. All teachers were selected based on recommendations forwarded to program administration, experience in test preparation, and general interest. Although teachers were experienced, only one had been involved in prior test preparation programs. Consequently, they were not extremely knowledgeable about coaching the PSAT. One teacher aide was also provided at each site.

Program Description

The 75-hour summer program included approximately 60 hours of test preparation and was divided in three content areas: (1) language arts, (2) mathematics, and (3) guidance. It was organized around the content of the actual PSAT with special emphasis on test-taking strategies, testwiseness, and drill. Three main textbooks were used: (1) Barron's PSAT/NMQST Guide, (2) Strategies for Taking SAT Test, and (3) Gruber's Inside Strategies for the SAT. At each site the program was divided into three groups of three classes on the basis of math PSAT pretest scores. Each site subgroup rotated among subject areas on a daily basis. Each class period was 50 minutes in length.

In language arts classes, students examined and reviewed vocabulary, sentence completion, verbal analogies, and reading comprehension sections of the PSAT. In the mathematics section of the course, basic mathematics, algebra, and geometry were emphasized. The guidance area explored awareness and understanding of test-taking strategies, values clarification, and career choices. Future plans and goals were also discussed in this component.

With the exception of the primary textbooks, there was no uniform curriculum for each component. Concentrating on a particular area, each teacher was allowed the flexibility of selecting his/her own manner of presentation and instruction. However, practice test-taking, drill, and testwiseness strategies were emphasized throughout.

Instruments

Student questionnaire. This 41-item self-report questionnaire assessed the content, instruction materials, and student understanding of testwise-ness principles of the program. Administered at the end of the program, survey items were coded on a four-point Likert Scale and were composed of items about general program assessment (i.e., "I learned test-taking strategies that I did not know before"), the effectiveness of materials ("How effective was Inside Strategies for the SAT?"), and understanding specific testwiseness principles ("In this program I learned that I should usually guess when I don't know the answer to a problem.") The higher an item was rated the more positive the particular dimension was assessed.

Teacher questionnaire. A 9-item multiple choice and open-ended questionnaire was completed by program teachers which asked them to document skills taught in the program, instructional strategies and evaluation criteria used, and problems encountered in implementation. Teachers also assessed the effectiveness of the instruction materials used in the program as well as its strengths and weaknesses. Recommendations for future test preparation programs were also solicited.

Classroom observations. To assess the implementation of the test preparation program, classroom observations were conducted at the three sites by two staff members from the Department of Research and Evaluation. Each subject area at the sites was visited once at the beginning of the program with the exception of one subject, which was observed twice during two different class periods. The subject areas of the double observations were different at each site so the number of total observations within each subject area were equivalent. These descriptive observations documented various

classroom variables including (1) the compatibility of the classroom activities and the program outline, (2) student time-on-task, (3) type and length of classroom activities, (4) materials used, and (5) the classroom learning climate. At the completion of each observation, brief interviews were completed with teachers that encompassed reactions to the materials, student reactions to the content, and the ability of students.

The PSAT. To determine student academic growth of program students, past PSATs were used, specifically forms S and T of the Fall 1985 PSAT. It should be noted that Lane students took Form 2 of the 1985 PSAT as the posttest. These tests were administered in alternate forms at the beginning and end of the program to both summer program and comparison groups. In addition, to assess the stability of PSAT scores, pre- and posttest scores were compared to results of the October 1986 PSAT. The PSAT is a standardized achievement test given to high school sophomores in preparation for the SAT. According to the College Entrance Examination Board (CEEB) (1985), the sponsor and governing body of the PSAT and SAT, the PSAT is "a multiple-choice test that measures developed verbal and mathematical reasoning abilities important for academic performance in college. It assesses ability to reason with facts and concepts rather than the ability to recall and recite them." Items assess reading comprehension, word meaning, sentence completion, basic arithmetic, algebra, and geometry. It is highly correlated with the SAT. There are 115 questions on the test, 65 in verbal and 50 in the mathematics section. Raw scores are corrected for guessing and converted to linearly derived scaled scores ranging from 20 to 80 with a mean of 50. Selection index scores are used for merit scholar selection and are computed as twice the verbal score plus the math score. Testing time is one hour and 40 minutes.

Procedure

Eligible students were sent letters in May 1986 inviting them to participate in the program. Although many eligible students did not receive these letters, those who enrolled in the six-week program were assigned to one of three sites, depending on which site was closest to their home or regular school. Students at each site were divided into three groups on the basis of their pretest scores and subsequently rotated among three teachers, one in each of the three content areas of language arts, mathematics, and guidance. On June 16, 1986, second letters were sent to eligible students at Lane, Lindblom, Curie, Kenwood, and Young inviting them to take two practice tests as comparison group students at either Lane or Curie. This was practical for many students who did not have time to enroll in the program. All students, summer program or comparison group, were administered alternate-form pre- and posttest PSATs before and after the program. Additional background information was also obtained. During the program, student and teacher questionnaires were completed by participants and classroom observations were conducted. October 1986 PSAT scores were also obtained for both groups.

Results

Program Implementation

The implementation of the TPPGTS was assessed by classroom observation to ascertain the degree to which it was enacted. Based on nine observations conducted for the program, three at each site, it appeared the program was implemented as intended. Observational records indicated that the vast majority of students participated in classroom activities and were on-task during class. In fact, all students observed during two classroom time intervals were judged as being on-task by observers, although it was noted that students were not always actively participating in classroom activities.

As intended, instructional activities were diverse and comprehensive. For example, during the 456 minutes of actual class time observed (approximately 51 minutes per observation), 36% of the instructional activities were devoted to independent or small group completion of assignments, 32% for listening to lectures or demonstrations, 23% for classroom discussion and recitation in response to teacher questions, and 9% for other group-related activities such as correcting answers to test questions. In addition, completing assignments independently or in small groups was the most frequent instructional activity as six of the nine observations noted.

The instructional activities observed consistently agreed with the outline and goals of the course. English classes primarily included review of vocabulary words and their root meanings and test practice. A variety of supplemental materials was also used to satisfy this objective including a word study guide, How to Ace the SAT, and The College Prep Game.

In the math section, lecture, classroom discussion, and practice test-taking were primarily employed. Teachers commonly explained the rationale behind correctly answering geometry and algebra items. In addition to the primary texts, other materials included a test preparation workbook, Mastering the SAT and past PSAT math exams. For example, the math instructor at Lane employed test practice and drill techniques from old PSATs to familiarize students with the PSAT.

Guidance component activities appeared to fall into two categories, (1) reviewing test-taking strategies and (2) exploring values, college, and career choices. While reviewing test-taking strategies was most often handled through lecture and class discussion, the latter activities were instituted as group and class discussion activities. For instance, during one class students filled out and discussed a career interest inventory. Following this activity, a job-simulation exercise was completed where student pairs reacted to possible job situations. Supplemental materials guided these group activities. Student progress was assessed in a variety of formal and informal methods including verbal feedback, monitoring, discussion, and practice test results.

Classroom observers also assessed the learning climate of the program. Results showed that on a four-point scale, student behavior was adequately controlled in the classroom (3.8), students were task-oriented throughout

class (3.4), students participated in class (3.6), and teachers facilitated classroom learning (3.6). Classroom observers also noted that the diversity of the instructional format stimulated student involvement (3.7), and the primary instructional materials were not consistently used during class (2.3). This latter finding suggests the lack of a uniform instructional focus.

Informal teacher interviews were completed at the end of each observation and indicated that level of interest in the program was influenced by the subgroup in which the students were placed. Teachers reported that motivation and interest level declined for the lower ability subgroups. Some teachers reported that the implementation of the program was negatively influenced by the lack of regular student attendance and the lack of math preparation for many students which changed the program from a review/test-taking focus to an instructional one.

Thus, TPPGTS appeared to be implemented as designed, as nearly all instructional activities and materials were used in accordance with the outline of the course. However, it should be noted that observations were few and conducted early in the program. To gage a comprehensive implementation of the program, additional observations should be conducted toward the end of the program. In addition, as has been noted in previous reports, many of the guidance activities were independent of test preparation such as values clarification, college, and career information activities. In fact, as the guidance teachers indicated, less than one-half of all class time was devoted to test preparation activities. Given that the design of this program is obligated to test preparation, these guidance activities are clearly incompatible with the program and may be unexpected by participating students. To rectify this state of affairs, at least the total content of the program should be reflected in its title, and at best, the guidance component should be purged of all such "nontest-taking" activities and incorporated with the other components.

Test Results

In order to assess the central question of the evaluation, that of whether summer program students made greater gains from pre- to posttest than the comparison group, a variety of statistical analyses was employed. To determine general within-group program effects, a paired-sample correlated t-test was used. This test is regarded as the most powerful for a paired, correlated sample (Hays, 1981). Second, to determine general PSAT program effects between summer program and comparison group students, Hotelling's (1931) T^2 statistic was computed across sites for verbal and math subtests. This multivariate statistic is also widely regarded as the most powerful test for a p-variate, simultaneous comparison (Marascuilo & Levin, 1983). Third, to determine substantive program effects, univariate analysis of covariance was used for the verbal and math subtest. Although multivariate analysis of covariance is generally more efficient, the verbal and mathematics subtests are considered separate and independent tests and will be analyzed along this line.

As shown in Table 3, paired t-test results indicated that summer program students made significant pre- to posttest gains on the verbal ($t=2.52$, $p<.05$) and math ($t=7.82$, $p<.001$) subtests. Their math subtest gain score (in points) of 5.2 (48.8 to 54.0) was over 3 times greater than their verbal subtest gain

Table 3

Mean Practice PSAT Pre- and Posttest Scores by Group and Site[1]

Site/ Occasion	VERBAL		MATH		SELECTION INDEX[2]	
	Summer Program. (St. Dev)	Comparison (St. Dev)	Summer Program (St. Dev)	Comparison (St. Dev)	Summer Pro. (St. Dev)	Comparison (St. Dev)
LANE (N=53/35)						
Pretest	42.0 (8.9)	43.9 (6.1)	48.7 (9.0)	50.4 (7.6)	132.6 (22.2)	138.3 (17.2)
Posttest	43.3 (8.5)	46.5 (7.7)	56.1 (8.2)	50.7 (7.2)	142.6 (21.9)	143.6) (19.1)
Gain	+1.3	+2.6**	+7.4***	+0.3	+10.0***	+5.3**
LINDBLOM (n=19/-)						
Pretest	43.7 (7.0)	- -	49.0 (6.2)	- -	136.5 (15.4)	- -
Posttest	46.2 (5.5)	- -	52.3 (6.1)	- -	144.6 (14.7)	- -
Gain	+2.5	-	+3.3*	-	+8.1**	
CURIE (n=21/12)						
Pretest	44.0 (7.5)	45.4 (8.1)	48.9 (7.2)	51.2 (7.8)	136.9 (15.3)	142.0 (20.3)
Posttest	45.1 (6.7)	45.3 (8.6)	50.3 (5.5)	52.8 (7.1)	140.4 (15.2)	143.3 (19.0)
Gain	+1.1	-0.1	+1.4	+1.6	+3.5	+1.3
TOTAL (n=93/47)						
Pretest	42.8 (8.2)	44.3 (6.6)	48.8 (8.1)	50.6 (7.6)	134.4 (19.5)	139.3 (17.9)
Posttest	44.3 (7.6)	46.2 (7.8)	54.0 (7.6)	51.2 (7.2)	142.5 (19.2)	143.6 (18.9)
Gain	+1.5*	+1.9*	+5.2***	+0.6	+8.1***	4.3*

[1] Scores used are scaled scores and range from 20 to 80.
[2] Selection Index (2*verbal + math)

* p<.05
** p<.01
*** p<.001

score of 1.5 (42.8 to 44.3). Although the comparison group, suprisingly, had a higher pre- to posttest verbal gain then the summer program group (1.9 points) ($t=2.30$, $p<.05$), their math gain score of 0.6 was not statistically significant. Thus, while both groups gained similarly on the verbal test, the summer program group made substantial gains on the math subtest over and above the comparison group. Mean total selection index (2*Verbal + Math) gain scores were also considerably higher for the summer program group (8.1 to 4.3 points).

Table 3 also indicates that pretest comparison group scores are higher than the summer program group for verbal, math, and selection index scores. However, independent t-test results indicated these average pretest scores were not statistically higher than summer program students. This further confirms the relative equivalence of groups before the onset of the program.

Table 3 also provides a PSAT score breakdown by program site. Results indicate that Lindblom program students gained 2.5 points on the verbal subtest while program students elsewhere gained approximately one point. Comparison group students gained very little on the verbal subtest with the exception of those tested at Lane, who scored 1.3 points higher than the summer program group. They were completely responsible for the gain of the entire group. At Curie, comparison group students had a slightly negative gain score from pre- to posttest test while the summer program group gained about 1 point. There was no comparison group at Lindblom.

In regard to the math subtest, Lane program students were primarily responsible for the total group gain as they improved 7.4 points from pre- to posttest. Lindblom and Curie students gained 3.3 and 1.4 points, respectively. Math gains for comparison group students were minimal, but suprisingly, comparison group students at Curie had a higher gain score than their summer program counterparts (1.6 to 1.4 points). Selection index score gains were also highest for Lane program students as they gained 10 points from pre- to posttest, twice the gain of their comparison group counterparts. Lindblom students, who did not have comparison group counterparts, gained 3.1 points on their selection index scores, similar to the total group average. Curie program students had, in addition to the lowest verbal and math gain scores, the lowest selection index gain scores (4.5 points). However, this gain was over three times higher than their comparison group counterparts. It should be noted that individual site results should be regarded with caution, especially at Lindblom and Curie, because of relatively small sample sizes. Despite this caution, it appears the program had a differential impact across sites as both teachers and content were different. These differences should be kept in mind when interpreting program effects.

Given the statistically significant gain scores of summer program students, between group analyses were conducted to determine substantive program effects. Results indicated a significant overall multivariate difference on verbal and math scores between summer program and comparison group students ($T^2=15.70$, $p<.001$). However, there were no significant differences between groups on the PSAT verbal subtest after accounting for verbal and math pretest scores. In fact, the comparison group had a larger gain score than the summer program group (1.9 to 1.5 points). However, significant differences between groups on the PSAT math subtest remained even after controlling for other background variables such as TAP reading, math,

and total scores, number of math and English courses taken in high school, and overall high school math and English grades. Given that comparison group students scored generally higher on the TAP, differences between groups became even larger after analyses of covariance. Group differences on PSAT verbal posttest scores remained nonsignificant after all analyses of covariance.

As a result of the above analyses, an estimation of program effects can be made on both the verbal and math PSAT subtests. This estimation, via analysis of covariance with multiple covariates, adjusts for pre-existing group differences on all included variables and generally increases the precision of estimates of program effects (Cook & Campbell, 1979). Independent subtest results indicated a positive treatment effect of 4.2 PSAT points on the math subtest after adjusting for group differences on the math practice pretest, 1985 TAP math scores, and 1985 TAP total scores. In other words, if the two groups started with the same math PSAT subtest, TAP math, and TAP total scores; the summer program group would have significantly out performed the comparison group by 4.2 points on the PSAT math posttest. Unfortunately, this positive treatment effect cannot be generalized to the verbal subtest. In fact, results indicate a negative treatment effect on the PSAT verbal subtest. That is, after accounting for initial group differences on the math pretest, TAP math subtest, and TAP total test, the comparison group out performed the summer program group by approximately 1 PSAT point. This difference was not significant. Thus, the verbal PSAT coaching had no apparent effect in raising participating students' test scores.

It should be noted that many other variables were included in the analysis of covariance model in order to adjust for pre-existing selection difference in the groups including number of math and English courses taken; grades received; specific subscales on the TAP; and combination of TAP, math, and verbal pretest scores. These added variables did not improve the precision of treatment effect estimates or the prediction of verbal and math posttest scores.

PSAT Scores by Stanine Group

Table 4 provides a breakdown of PSAT scores by reading and math stanine groups. Employing paired t-tests, there were significant pre- to posttest gain score differences for nearly all TAP reading and math stanine groups. As is also shown, there are clear divisions in test scores between stanine groups. Each stanine group received higher PSAT subtest scores than the preceding group. Referring to the TAP reading stanine breakdown, ninth stanine students made the only significant pre-posttest gain on the verbal subtest, although stanine seven students also had a nearly identical gain. In regard to pre- and posttest math scores, students in all reading stanines made highly significant gains but ninth stanine students obtained the highest math gain score (5.7 points). Ninth stanine students also obtained the highest average selection index score (10.1 points).

A similar pattern emerged in comparing math PSAT scores by math TAP scores. Again, ninth stanine students made the only significant average verbal subtest gain score (3.7 points) while stanine seven and eight students' gain scores were less than one point. Regarding math subtest scores, math stanine seven students obtained the highest average gain score (6.7 points),

Table 4

Mean Practice PSAT Scores by TAP Reading and Math Stanine Groups
(N=93)

	TAP READING STANINES			TAP MATHEMATICS STANINES			R/M-99[1]
Test/Occasion	7 (n=31)	8 (n=37)	9 (n=25)	7 (n=37)	8 (n=34)	9 (n=22)	99 (n=10)
Verbal Pre	38.2	44.3	46.3*	40.9	43.7	44.6	49.2
Verbal Post	40.5	44.6	48.5	41.8	44.6	48.3	51.7
Gain	2.3	0.3	2.2*	0.9	0.9	3.7**	2.5
Math Pre	48.6	46.8	52.0	43.1	50.1	56.4	59.3
Math Post	52.8	52.4	57.7	49.8	53.4	61.9	64.1
Gain	3.2***	5.6***	5.7***	6.7***	3.3**	5.5***	4.8*
SI[2] Pre	125.0	135.4	144.6	124.8	137.5	145.6	157.7
SI Post	133.8	141.6	154.7	133.3	142.6	158.5	167.5
Gain	8.8**	6.2**	10.1***	8.5***	5.1*	12.9***	9.8***

[1] This group scored in the ninth stanine in reading and math
[2] Selection index (2*verbal + math)

* p<.05
** p<.01
*** p<.001

slightly higher than ninth stanine students (5.5 points), and twice as high as eighth stanine students (3.3 points). However, the large actual score differences between stanine groups should not be forgotten as math subtest scores increased by six to seven points for each stanine level. Similarly, math stanine nine students also received the highest selection index average gain score (12.9 points), although all stanine groups made significant pre-posttest gains. Eighth stanine students obtained the lowest selection index score gain (5.1 points). Thus, the data generally indicate fairly high gain scores for seventh stanine students but then dropping somewhat for eighth stanine students and then rising even higher for ninth stanine students.

Of additional interest are PSAT results of students in the ninth reading and math stanines. As shown, they received the highest subtest scores of any other group but not the highest gain scores. Although their verbal gain scores were nigher than nearly all other groups (2.5 points), it was not statistically significant. Math (4.8 points) and selection index (9.8 points) gain scores were statistically significant although they were not quite as high as those of TAP reading or math ninth stanine students. It should be noted that the relatively small sample size of this group limits the above results.

Thus, ninth stanine TAP reading and math students appear to take the most advantage of the program as their gain scores are generally higher than other groups. Given the goal of producing more National Merit scholars, ninth stanine math and reading students are the most likely to satisfy such a goal, witness that their selection index scores are at least 9 points higher than any other stanine group.

Analysis within the Summer Program Group

Also of interest in this investigation were differences in PSAT performance by various student subgroups including gender, PSAT experience, and reasons for enrolling in the course. In regard to gender differences, male and female program students scored similarly on pre- and posttests as there were no statistically significant differences on verbal or math subtests. Male and female students had identical verbal pretest scores (42.8) and similar verbal posttest scores (43.7 and 44.8), respectively, while male students had slightly higher math pretest (50.0 to 47.4) and posttest scores (54.0 to 53.6). With the exception of the written expression TAP subtest, whereby female program students scored significantly higher than their male counterparts, groups were similar on all background variables such as year in school and TAP reading and math subtests.

PSAT experience appeared to have a greater effect on scores than gender. Program students who had taken the PSAT before, pretested significantly higher on the verbal ($t=2.60$, $p<.01$) and math ($t=4.19$, $p<.0001$) subtests of the PSAT than students who had not taken the test previously. This gap was narrowed by the posttest as there were no significant differences between groups on the verbal subtest and smaller differences on the math subtest ($t=2.19$, $p<.03$). However, students who had previous PSAT experience were a higher achieving and older group than novice test-takers as they scored significantly higher than other students on the TAP math subtest, the total TAP, and the TAP basic subtest. They also were significantly older and had taken more math courses than their PSAT counterparts.

It was also of interest to compare test scores by the nature of participation in the program. For example, it was hypothesized that students who enrolled in the program specifically to improve their PSAT scores would have greater gains than other students, such as those who entered the program only because their parents expected them to or thought it would be a good way to meet other students. Results indicated no significant differences in verbal or math subtest scores between students who enrolled only because their parents expected them to and those that enrolled for other reasons. There were also no differences between students who enrolled because they wanted to do better on tests and those who enrolled because they didn't have anything else planned for the summer or thought it would be a good way to meet people.

Supplemental Reports of Program Effects

To supplement pre- posttest data, student and teacher questionnaires were completed during the last week of the program.

Student questionnaire. Results of the 101 student questionnaires, not necessarily including those in the matched program group, indicated that they rated the program quite positively on a scale from one to four, with four

being the most positive. For example, of the 22 items assessing general program content, students thought the English (3.4), math (3.4), and guidance components were helpful in preparing for the PSAT. Their vocabulary (3.4) and problem-solving (3.2) skills were improved, and as a result of the program they were better prepared to take the October 1986 PSAT (3.4).

Not so positively assessed features of the program were the relevance of the guidance component (2.8), newness of the materials (2.5), and meeting everyday (2.7). Although students were divided, they indicated that taking the practice tests was somewhat more effective than class instruction in preparing for the PSAT (2.7).

A second dimension assessed by the questionnaire was the effectiveness of the instruction materials. Students, on the average, indicated the instruction materials were effective (3.5) and rated the primary instruction materials in the following order: Barron's How to Prepare for the PSAT/NMSQT (3.4), Barron's Strategies for the SAT (3.1), and Inside Strategies for the SAT (3.0). It should be noted that individual copies of these materials were not provided so they are assessed only in terms of their classroom use.

A third function of the questionnaire was to ascertain the kinds of test-taking strategies learned in the program. Students responded to a series of 12 testwiseness items and rated them on a scale from one to four. Students agreed that many concepts were learned in the program including short-cuts to doing PSAT problems (3.4), using time wisely (3.5), the importance of knowing algebraic and geometric operations, and understanding how the PSAT is designed (3.4). More specifically, students stated their level of agreement with a number of principles taught in Inside Strategies for the SAT. Eighty-three (83) percent correctly disagreed or strongly disagreed "that the only way to do well on the vocabulary test is to memorize as many words as possible." Only 25% of the students strongly disagreed. In reference to the question regarding reading directions to the PSAT if already known, 70% of the students correctly disagreed or strongly disagreed that directions should be read in this instance. More convincingly, 96% of the students agreed or strongly agreed that test choices should be tried in reverse order when the answer to an item is not known. In response to a item about guessing, 70% of program students agreed or strongly agreed that they should usually guess when they don't know the answer to a problem. However, too many students disagreed (30%), suggesting they would not usually guess in this instance.

Unfortunately, all item responses did not indicate students learned what they read or were taught. For example, 60% of the students agreed that they should first read the questions following the passage before reading it, although Gruber recommended in Inside strategies for the SAT that the passage should be read first.

Questionnaire responses were also compared between Lane site students and those at the other sites in order to help explain gain score differences between sites. The only item found to be significantly higher in favor of Lane students regarded guessing. Lane students learned to a greater degree than others that they should usually guess when they come to a item for which they didn't know the answer ($t=4.57$, $p<.0001$). Thus, the higher math gain scores of Lane students was not entirely explained through questionnaire responses.

Thus, for the most part, students learned important test-taking principles that will help them on the PSAT, although in some cases, such as reading directions and guessing, a substantial number of students misconstrued proper testwiseness principles.

Teacher questionnaire. Teacher survey responses reinforced the diversity of instruction materials and strategies used in the program. In addition to using the assigned materials, teachers indicated they supplemented instruction with test preparation materials, college guidebooks, and mathematics and English textbooks. These materials were employed for a variety of instruction activities including lecture, discussion, demonstration, independent work, and innovative activities (i.e., games and group exercises). The frequency of these activities (measured on a scale from one to four with four indicating daily use) varied considerably as discussion (3.5), independent seatwork (3.0), and oral recitation (2.8) were used more frequently than demonstrations (2.5) and lectures and presentations (1.2). Corresponding abilities taught with these materials and activities were primarily test-taking skills, listening, reading, study skills, stress management, and problem solving.

However, the use and frequency of such materials and activities varied by component. Not surprisingly, English class activities emphasized discussion, oral recitation, and seatwork with the objective being vocabulary and reading development. Supplemental vocabulary handouts and practice exercises were used toward this end. The mathematics component centered on test practice and drill toward the goal of improving problem-solving, computation, and test-taking skills. Supplemental materials such as math workbooks were also used. For example, at Lane the math instructor administered past PSATs every week and provided feedback after each test as the primary instruction focus. On the other hand, guidance component activities stressed test-taking study skills and other skills not specifically related to test preparation such as stress management, image building, and college and career planning. Instruction activities most frequently used were discussion, independent seatwork, and demonstration. A plethora of materials was employed in addition to test-taking materials and included counseling and college planning guides and career development manuals. Group exercises were also employed toward the goal of personal development (i.e., group dynamics, leadership skills, and values clarification).

Teachers also assessed the effectiveness of the primary instruction materials and reported on a scale from one to four that Barron's test preparation book was most effective (3.4), while the other primary source book Inside Strategies for the SAT, was rated lower (2.7). However, teachers noted that the latter would have been more effective if copies were available to students (i.e., "would have been better if every student had a copy"). Barron's Strategies for Taking Tests and Mathematics for the College Boards were also rated positively, although some teachers did not rate their effectiveness.

Mastery of program objectives was assessed in a variety of ways but most frequently by in-class homework assignments, participation in class discussion, teacher-made tests, and oral recitation. For example, the English class at Curie had daily vocabulary quizzes to familiarize students with word meanings.

In addition to documenting the techniques and materials used in the program, teachers noted strengths and weaknesses of the program and made recommendations toward its improvement.

In regard to limitations of program implementation, teachers reported that regularity of student attendance (eight teachers so indicated), coordination between components and with the central office, and student interest in class presented problems in the program. They also voiced concern over planning aspects of the program including criteria for student selection, recruitment procedures, ability of students recruited, and the quality of the teacher inservice. In regard to the criteria for selection, one teacher noted the importance of the "selection of students with high G.P.A.'s as well as high test scores...It takes a combination to be successful in entering college."

In contrast, strengths of the program noted by teachers included its positive emphasis on test-taking strategies, the small class size, and the general concept of test preparation. For example, one teacher indicated "The concept is very good. The end results were very promising" while another stated the "Small classes added intimacy and allowed for personal interaction."

Recommendations made about the program were numerous and included improving the criteria of selection by using grades and teacher recommendations, shortening the program to an after-school program, recruiting students much earlier in the spring, providing more teacher in-service sessions, improving the coordination with the central office, and giving students academic credit for participating. Most of these recommendations have been made in earlier reports and must be resolved before another program is implemented.

October 1986 Results

To determine the accuracy and stability of PSAT practice tests, results of the October 1986 PSAT were solicited for summer program and comparison group students. These results as well as matched pre- and posttest results are listed in Table 5. As can be seen, summer program students had a mean verbal scaled score of 46.5 compared to 48.4 for comparison group students. These scores indicate posttest-to-October 1986 test score gains of 2.1 and 1.8 points, respectively, slightly higher than pre-posttest gain scores. Summer program (54.1) and comparison group (53.1) mean math scaled scores represented 0.8 and 1.6 point gains from the posttest. This general upward movement of test scores is not consistent with 1985 findings and suggests, especially for the verbal subtest, that regular fall academic school work helped improve test scores more than the program or that motivation to do well was more prevalent on the October test. The small sample sizes obtained should not be forgotten.

Table 5

Mean PSAT Scores of Students taking the October 1986 Test by Group

Test/ Group		June 1986 (Pretest)	July 1986 (Posttest)	October 1986 (Actual)	July to October Gain
Summer Program (n=72)	Verbal	43.0 (7.9)	44.4 (6.9)	46.5 (8.2)	2.1
	Math	49.3 (7.6)	53.3 (7.1)	54.1 (7.7)	0.8
Comparison (n=37)	Verbal	44.9 (6.1)	45.6 (6.8)	48.4 (6.8)	1.8
	Math	51.3 (7.3)	51.5 (7.2)	53.1 (6.4)	1.6

Discussion

The major evaluation question addressed was whether summer program students outperformed comparison group students from pre- to posttest and if this performance can be attributed to the program. Results indicated a significant positive program effect for the PSAT math subtest after accounting for PSAT math pretest scores and TAP math and total scores. This was especially apparent at Lane where program students gained 7.1 PSAT points over a comparison group. Relatively small sample sizes precluded interpretation of program effects at the other sites.

In regard to the PSAT verbal subtest, there were no overall significant differences between summer program and comparison groups on the PSAT verbal posttest after accounting for differences on the verbal pretest, TAP reading and total test scores, high school English grades, number of high school English courses taken, and PSAT math pretest scores. In fact, results showed an overall negative program effect for the verbal subtest as comparison group students obtained higher gain scores from pre- to posttest than summer program students, especially at the Lane program site. The estimated overall program effect from analysis of covariance was -.7 PSAT verbal score points or a loss by the summer program group relative to the comparison group or approximately 7 points on the SAT scale. Although this loss is statistically nonsignificant, it does indicate the effectiveness of the verbal coaching component of the program was poor.

The inconsistent results obtained across sites suggest that the program was implemented differently at each site. For example, at the Lane site the instructor emphasized systematic practice-testing and feedback with past PSATs that was not apparent at the other sites. Differences in teacher experience and knowledge of test preparation may have also played a role. The small sample sizes at the non-Lane sites may have further exacerbated observed results.

While these different results should be kept in mind, it should not dissuade us from interpreting verbal and math coaching effects. The primary question to be addressed here is the validity of attributing the positive and negative program effects to the test preparation program.

Establishing the Validity of Observed Program Effects

The nonequivalent control group design with pretest and posttest measures used in this study is a generally interpretable design for establishing the internal validity of observed results. Results are internally valid when program effects can be attributed directly to the program and alternative explanations of program effects are ruled out. The present design typically rules out many threats to interval and external validity that may bias estimates of program effects. Following Cook & Campbell's (1979) method of determining the internal and external validity of observed program effects, the primary validity threats are tested individually for plausibility so that the nature of test score gains can be found. Elimination of all or most threats to internal validity lends support to the attribution of test score gains to program effects while elimination of external validity threats would support the representativeness of present findings to other student populations and settings. The primary concern here will be the observed PSAT math coaching effects, especially at Lane, since the positive evaluation results were observed here. However, interpretation of no program effects for the PSAT verbal coaching component will be discussed separately. The discussion will begin with internal validity threats.

Threats to Internal Validity Ruled Out

Based on the design and results of the study, the following threats may be ruled out:

History. This threat occurs when criterion scores of an experimental group are influenced by forces outside the context of the program such as other people or other instruction. The present control group design generally accounts for this threat as it is assumed outside forces are influencing both treatment and comparison groups equally. This potential effect would then be controlled in the pre- posttest data. History is a special concern in education because students are continuously exposed to an amalgamation of instruction programs, all of which may influence each other. In the present study, this threat is further neutralized by the relative short duration and isolated conditions (summer school) of the program. History effects usually occur most frequently with longer running programs.

Maturation. This threat is of concern when an observed treatment effect may be due to the general maturing process (i.e., growing older, wiser, or more experienced) rather than the program. As with history, maturation effects usually take place over a long period of time and would rarely occur over a six-week period. Further, the present design generally controls for maturation effects, and it is unlikely differential effects were operating between groups.

Testing. Testing effects may occur when observed program effects are the result of participants becoming more familiar with a criterion test such as when the same test is used for all testing occasions. Again, the use of a

pre-posttest control group design usually accounts for this potential rival hypothesis, because both groups would take advantage of such a situation. In the present study, this threat is further mediated by the use of alternative test forms from pre- to posttest.

Instrumentation. This threat typically occurs when the measuring instrument is changed in some way between the pre- and posttest or when groups exhibit "floor" or "ceiling" effects. Both effects appear to be controlled in this study. The alternate-form PSATs used produce identical linear standard scores that are statistically equated. "Floor" or "ceiling" effects were not an issue because mean test scores hovered around the middle range of the PSAT scale.

Statistical regression. This threat generally refers to the upward movement of a pretested experimental group to its population mean at the posttest which is mistaken for a treatment effect rather than a statistical artifact. The threat is most ominous when the treatment group has a much lower pretest score than the comparison group and criterion measures are unreliable. The latter condition is not plausible since the PSAT is a nationally standardized test that has high test-retest reliability. The former concern is also minimized because (1) the summer program and comparison groups did not perform signficantly differently on the math pretest to render regression a major threat and (2) the pattern of results obtained on the math test in which the lower-scoring summer program group overtook the comparison group in a cross-over fashion by the posttest. As explained by Cook & Campbell (1979), this outcome reduces the likelihood of a regression alternative explanation because it is not reasonable to expect the summer program group to regress above the posttest score of the comparison group.

Mortality. Mortality is a threat to observed program effects when a substantial number of students drop out of the program after the pretest and it is found that those students who drop out have different characteristics than students who stayed in the program. When attrition is high, sample representativeness is compromised and estimated treatment effects become biased. Although evaluation data were reduced to include only those students who completed pre- and posttests, consequently eliminating this validity threat, substantial reduction of the program sample is problematic. Sixty-four (64) percent of those students pretested participated to some degree in the program and took the posttest, a fairly positive retention rate. Of the comparison group, only 38% completed both pre- and posttests, indicating the interest and motivation of this reduced group in improving their test scores. However, it was found that those students who dropped out of the program after the pretest received significantly lower verbal and math subtest scores than the final summer program group.

Other major threats to internal validity ruled out. Other plausible alternative hypotheses to observed PSAT math coaching effects ruled out in this study include (1) diffusion of treatment, (2) compensatory equalization of treatments, (3) compensatory rivalry, and (4) resentful demoralization. These threats are used usually to explain minimal or no program effects and involve contamination between experimental and comparison groups. Thus, they are not directly applicable to math coaching results. The most likely threat in this study would be resentful demoralization, whereby the comparison group reacts negatively to its no treatment status and deliberately lowers its test

performance. This would have the effect of inflating the estimated positive program effect. However, given the exclusively voluntary nature of the comparison group, it is unlikely these students resented their no treatment status.

Threats Not Ruled Out as Explanations for Program Effects

Selection. Selection is a threat and potential explanation for observed program effects when there are preexisting differences between experimental and comparison groups that cannot be measured or controlled. It is especially a problem when recruitment procedures of groups are different and assignment of participants to experimental and comparison groups is not controlled. Although the use of analysis of covariance in the present study controlled the effects of some measured characteristics between groups such as practice pretests and achievement test scores, it cannot control for unmeasured characteristics that may be different between groups. These other unmeasured characteristics may have influenced the positive math coaching effects obtained.

Interest in improving PSAT scores and motivation to do well appear to be the primary threats as a result of selection. The fact that summer program students participated in an in-depth test preparation program suggests they were more interested and motivated in improving their scores than the comparison group. The magnitude of this effect is unknown, but it surely must be considered in interpreting the validity of the observed math coaching effect. Fortunately, though, this motivation/interest effect is minimized by two findings. First, comparison group students were obtained on the basis of their interest in improving their own PSAT scores. The choice to take two practice tests was completely voluntary and indicated they were also interested in improving their test scores. However, many students did not have time to enroll in the program or did not receive a letter inviting them to participate. It was not because they were not interested. In addition, nearly all comparison group students indicated they would be taking the October 1986 test.

Second, the plausible explanation that summer program students were more interested in improving their test scores may be offset by the fact that the comparison group was a higher achieving group than the summer program group. They scored higher on all achievement test measures and pretests. This information should also be taken into account.

Thus, the probability of the observed math coaching effect being the result of selection differences is reduced by the explicit interest of the comparison group in improving its PSAT scores and the possibility that it is a higher achieving group. The magnitude of this effect remains uncertain, and if this threat exists at all, it is most likely a modest one, but it cannot be ruled out.

Interactions with selection. These threats occur when selection differences cannot be ruled out and resulting differences may be combining with other internal validity threates such as maturation, testing, and history. These threats are difficult to estimate for non-equivalent groups because all selection differences cannot be measured or obtained. Selection-instrumentation can be eliminated quickly as both groups scored at approximately equal intervals on the PSAT, thus, results could be interpreted

similarly. Selection-maturation effects are minimized by the relatively short duration of the program and the finding that within-group variances decreased from the pretest to the posttest for both groups. This suggests that the selection-maturation threat was at best minimal. In the latter case it is assumed that if selection-maturation is operating, then differential growth between groups should be occurring within groups as well, and within-group variances do not indicate this occurrence (Cook & Campbell, 1979).

However, two other interaction threats cannot be dismissed so easily--selection-history and selection-testing. Selection-history, the most serious threat, occurs most typically in disseminated treatment programs where sites receive different kinds of instruction, and program effects are concentrated at one particular site. Such is the case with the math coaching component, since the primary treatment effect was concentrated at Lane. One must ask if there was some specific event or local history that enabled students at this site to gain over 7 points. The other two sites, although they were composed of much smaller numbers of students showed gains of only 3.3 and 1.4 points, respectively. In addition, all three groups started with nearly identical math pretest scores. Taking out math PSAT scores at the Lane site reduces the average gain score from 5.2 to 2.3 and the unadjusted program effect from 4.6 to .7 points.

Thus, the apparent site-specific math PSAT program effect indicates that specific content information at Lane may have been responsible for the major program effect and not the program in general. As previously noted, teacher characteristics may have also played a part in observed test score differences between sites. Although the teachers recruited for the program were experienced in their particular subject areas, the Lane site instructor may have been more adept in preparing students for the PSAT.

Selection-testing. An ironic feature about this evaluation is that the effectiveness of a test preparation program is determined by tests. Thus, the summer program students are exposed to more test practice and past PSAT tests than their comparison group counterparts. This increased exposure to tests per se, rather than the instruction of the program may have been responsible for observed program effects. Although it can be argued that test practice and completion of former PSATs are an intimate part of the program instruction, determination of which component is the most effective in improving test scores is a highly relevant issue, and cannot be resolved in this study. Self-reported student questionnaire data indicated that 30% of those surveyed indicated that the tests were more important than the instruction in learning about the SAT.

The Status of the Verbal Coaching Component

Results indicated a small but negative program effect on the verbal PSAT from pre- to posttest. Thus. the verbal component may have been a detriment to summer program students. However, the estimated loss was less than 1 PSAT point. This negative and unusual result is difficult to explain as internal validity threats are not applicable to this result. One possible explanation is selection. Enrollment in the program favored those students who had not made summer plans, thus it is probable the comparison group is more academically involved than their program counterparts. Consistently higher pre- and posttest and achievement test scores support the greater academic experience of this group.

However, considering that over 25 hours was devoted to the verbal component resulting in an average gain of 1.5 points, this must be viewed as a failure of the program and its administration, especially since the goal of the program is to produce national merit scholars from primarily above average students. In all fairness, it should be noted that the verbal subtest is less amenable to coaching effects than the math subtest.

Threats to External Validity

The purpose of a nonequivalent control group design is to eliminate most threats to internal validity and establish a program effect. It is not particularly well designed to produce externally valid results or results that would be similar across other students and settings. The three major threats to external validity described by Cook & Campbell (1979) are the (1) interaction of selection and treatment, (2) interaction of setting and treatment, and (3) reactive arrangements. The latter two threats appear to be ruled out in this study. The interaction of setting and treatment or treatment effects varying with the setting can be generally ruled out on the grounds that educational test preparation programs are intended for a homogenous setting, the classroom, and, all other things equal, would probably vary minimally across such environments. Also ruled out is reactive arrangements, the probability that participants in other educational settings will react differently to the program in ways that change the magnitude of program effects. This threat is greatest in laboratory studies where results are obtained in contrived and artificial settings. This is not the case with classroom research as testing and program instruction are regular features of education.

Unfortunately, the most relevant threat to be ruled out, selection-treatment, cannot be. This external validity threat limits the generalizability of results when program/treatment students are not representative of students-at-large. Obviously, program students are a selective student group in regard to their motivation to do well on the PSAT and as test-takers. Program students were in the upper-third of the TAP achievement test distribution; thus, results cannot be generalized to students testing lower than this restricted range. It should also be noted that coaching research has not considered restricted student groups in determining program effects.

Generalizability of criterion scores. While positive coaching effects were found on the math subtest, the nature of these effects and their meaningfulness are uncertain. As Anastasi (1981) has indicated, it is not clear whether test score gains due to coaching also result in improved criterion score performance. The major purpose of test preparation programs, especially on ETS tests, is to learn the test's structure, how it is designed, testwiseness strategies, and effective response patterns and not substantive content training. Thus, it is uncertain whether a student who scored 52 on the PSAT after a test preparation program would have the same performance capacity as someone who scored 52 without test preparation. The relatively small sample size of the study should also be considered, especially at the Curie and Lindblom sites. Generalizability of such results should be made with caution.

The Status of the Program

On the basis of the above discussion, it is possible to interpret the observed math coaching effect. Even though most threats to internal validity may be ruled out as influencing observed effects such as history, maturation, testing, mortality, instrumentation, and to a degree, selection, a general program effect did not exist. The primary positive effect of the program was observed at Lane where gain scores can be attributed to the systematic practice testing and analysis of past PSATs and possibly the teacher. Thus, the positive program effect is site-specific but substantial. The fact that students coached at Lane gained 7.1 PSAT or 71 SAT points over their comparison group strongly indicates the positive effect of practice testing, drill, and feedback. Small samples at the other two sites preclude interpretation of program effects.

These results indicate two major conclusions regarding this program: (1) the desirability of using former PSATs as instruction materials almost exclusively for practice and drill and (2) the need to establish clearer and more uniform guidelines for instruction. The use of past PSAT forms serves the relevant function of familiarizing students with actual PSAT items rather than approximated PSAT items, thus, giving students a greater sense of the type of items they can expect on the test.

In regard to instruction uniformity, it is imperative that future test preparation programs use a standardized curriculum for instruction, especially when the program is disseminated to multiple sites. The primary advantage of a standardized curriculum is that it minimizes teacher and content differences across sites. It also functions as a guide or "road map" for teachers to systematically follow and adhere to when they are not very familiar with the material. Also important is the selection and training of teachers familiar with the nuances of ETS-developed tests.

The status of the verbal coaching program is bleak. Results indicated a negative program effect rather than a positive one, although the magnitude is not statistically significant. While this result may have been affected by the nonequivalence of the groups, it is unlikely the effect was major. Thus, the content of the verbal coaching must be changed significantly. A more direct emphasis on test practice and drill with former PSATs is suggested. The current practice of using materials that have little relation to the PSAT (i.e., Barron's materials, English texts) should be eliminated in favor of consistent drill under actual testing conditions and analysis of responses. If the goal of coaching is to improve test scores only, then coaching should correspond as closely as possible to the content of the test. The public availability of past PSAT and SAT forms enables this compatibility to be high.

Comparability with Other Studies

The results of this study, although inconsistent across sites, support the positive effects of PSAT math coaching. The overall estimated math coaching effect of 4.2 PSAT points (42 SAT points or 71 points for Lane students) is much greater than the 10 to 15 point gains reported in the SAT review literature (DerSimonian & Laird, 1983; Slack & Porter, 1980), although negligible results were found at the other comparison group site. The present results do not conform to the logarithmic model of Messick and Jungeblut

(1983) as it was calculated that a math score gain of 40 points would require approximately 107 hours of instruction. Approximately 30 hours of math instruction was included in the program, indicating significantly greater time-effectiveness than other studies. This was especially apparent at Lane where summer program students gained 7.1 points over the comparison group.

The site-specific math program effect, then, does not undermine the positive coaching effects, although the nature of these effects is uncertain. These results are consistent with the 1984 and 1985 findings in which program students gained 4.4 and 3.5 PSAT points, respectively, from pre- to posttest (Chicago Public Schools, 1986). In regard to the verbal component, the present finding (1.5 PSAT gain score) is not consistent with the 1985 and 1984 results as it was found that program students gained 3.3 and 7.1 PSAT points, respectively. These differences as well as the site-to-site differences of 1986 may be explained, in part, by the nonstandardization of materials and teacher characteristics over the past three years.

Summary

TPPGTS was generally observed to be implemented as designed and satisfactory to both teacher and student participants. Teachers indicated the positive emphasis on test-taking strategies and small class size was facilitative of an effective program, while students noted they were better prepared to take the October 1986 PSAT. However, teachers suggested numerous changes be made such as shortening the program and changing the criteria for student selection.

Students also indicated that the guidance component was not compatible with test preparation and meeting everyday was cumbersome. Test results showed differentiated math and verbal subtest gain scores across sites, although verbal gain scores were negligible between comparison and summer program groups. A highly significant math program effect was found at Lane that completely accounted for the total group effect. It was found that the nature of the program at this site was different than the other two as systematic use of past PSATs was utilized.

Recommendations

Based on the above results, the following recommendations are made about TPPGTS:

1. Emphasize practice and drill with past PSATs. The tremendous math gain scores of the practice-test dominated instruction at Lane further supports the effectiveness of practice and review of past PSATs in improving test scores. Future programs should use this strategy in both verbal and math components. Again, as summarized by Anastasi (1981), the closer the correspondence between the program and the test situation, the higher will be the gain score. However, this also results in limited improvement of criterion behavior, that is doing well in course work and school in general.

2. Train teachers in the most effective ways to coach the test. Because effective teachers are essential to the success of any test preparation program, it is imperative they are trained in the intricacies of test coaching. Knowledge of how the test is developed, what skills are tested, and testwiseness principles are necessary for instructors to understand and teach. More extensive training sessions about the PSAT or SAT should be provided for the teachers.

3. Standardize the instruction and materials. Divergent results across sites may be attributed to the flexible use of materials and instruction methods by teachers. Teachers were able to structure their curriculum as they saw fit. However, since the usefulness of various instruction practices has been supported, those practices could be of top priority for implementation into the classroom. Thus, by structuring the program along the lines of practice and drill with past PSATs, the uniformity of the program may be established and the diversity of instruction minimized across sites.

4. Improve the student selection criteria. As has been discussed in previous reports, the exclusive reliance on TAP scores as the selection criteria undermines the identification of gifted students. Other criteria such as grades, teacher/counselor recommendations, and interest are also important for selecting students for the program. A consensus of teachers also indicated a problem with the selection criteria. However, altering the selection criteria will require more coordination between the schools and the central office. This will necessitate an earlier student selection process.

5. Refine the identification and selection of gifted students to be served by this program. Evaluation results indicate that ninth stanine students in reading and math outperformed other students on the PSAT and came closest to meeting the selection index cut off score. They scored, on the average, at least 9 points higher than any other stanine group. If the goal of the program is limited to increasing the number of National Merit scholars, ninth stanine students appear to have the best opportunity for achieving this objective.

6. Change the program to a general SAT/ACT preparation program. As discussed in the summary evaluation report (Chicago Public Schools, 1986), the present goal of increasing the number of National Merit scholars has not been satisfied as nearly all students who participate in the program are well below the PSAT selection index cut off score of approximately 200. A more realistic and influential program goal would be to focus on general score improvement on the primary college entrance examinations, the SAT, or the American College Test (ACT). Coaching for these tests would have a much greater impact on college entrance and could also provide more scholarship money for students to pay for college. Giving the general student population an opportunity to participate could increase the benefit of the program.

7. Give students an incentive for participating in the program. One way to increase the participation of gifted students and/or students in general is to offer some incentive such as academic credit, a certificate or letter of completion, or some such reinforcement that will, at least partially, improve the motivation of students to enroll and stay in the program. Many teachers also recommended this change in lieu of the fact that they indicated student attendance and motivation were problems.

8. Since coaching programs produce short-term gains and do not improve cognitive skills necessary for doing well in college, emphasis should be given to programs that develop long-term cognitive skills. The educational significance of improving college entrance examination scores may be limited to the test score itself. The SAT and PSAT measure a very limited set of abilities, and if overemphasized, downplay essential competency skills needed for learning. Coaching concentrates on testwiseness strategies and idiosyncratic qualities of tests rather than the development of cognitive skills. Rather than provide short-term gains on a test of questionable predictive validity, instructional programs on cognitive skills and problem-solving will provide the most effective foundation for academic success.

References

Anastasi, A. (1981). Coaching, test sophistication and developed abilities. American Psychologist, 36 (10), 1086-1093.

Chicago Public Schools (1986). A review of test preparation programs for gifted and talented sophomores, 1983-1985. Department of Research and Evaluation. Chicago, IL: Author.

College Entrance Examination Board (1985). A counselor's guide to helping students learn from the PSAT/NMSQT. Philadelphia, Pa: Author.

Cook, T.D. and Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin Company.

DerSimonian, R. and Laird, N. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. Harvard Educational Review, 53 (1), 1-15.

Dear, R.E. (1958). The effects of a program of intensive coaching on SAT scores (RB58.5). Princeton: Educational Testing Service.

Dyer, H.S. (1953). Does Coaching Help? College Board Review, 19, 331-335.

French, J.W. (1955). An answer to test coaching. College Board Review, 27, 5-7.

Hays, W.L. (1981). Statistics. New York: Holt, Rinehart and Winston.

Hotelling, H. (1931). The generalization of student's ratio. Annals of Mathematical Statistics, 2, 360-378.

Marascuilo, L.A. and Levin, J.R. (1983). Multivariate Statistics in the Social Sciences. Monterey, CA: Brooks/Cole.

Messick, S. and Jungeblut, A. (1981). Time and method in coaching for the SAT. Psychological Bulletin, 89 (2), 191-216.

Slack, W.V. and Porter, D. (1980). The scholastic aptitude test: A critical appraisal. Harvard Educational Review, 50, 154-175.